Depth-induced Multi-scale Recurrent Attention Network for Saliency Detection

Yongri Piao Wei Ji Jingjing Li Miao Zhang* Huchuan Lu
Dalian University of Technology, China
yrpiao@dlut.edu.cn, {jiwei521,lijingjing}@mail.dlut.edu.cn, {miaozhang,lhchuan}@dlut.edu.cn

Abstract

In this work, we propose a novel depth-induced multi-scale recurrent attention network for saliency detection. It achieves dramatic performance especially in complex scenarios. There are three main contributions of our network that are experimentally demonstrated to have significant practical merits. First, we design an effective depth refinement block using residual connections to fully extract and fuse multi-level paired complementary cues from RGB and depth streams. Second, depth cues with abundant spatial information are innovatively combined with multi-scale context features for accurately locating salient objects. Third, we boost our model's performance by a novel recurrent attention module inspired by Internal Generative Mechanism of human brain. This module can generate more accurate saliency results via comprehensively learning the internal semantic relation of the fused feature and progressively optimizing local details with memory-oriented scene understanding. In addition, we create a large scale RGB-D dataset containing more complex scenarios, which can contribute to comprehensively evaluating saliency models. Extensive experiments on six public datasets and ours demonstrate that our method can accurately identify salient objects and achieve consistently superior performance over 16 state-of-the-art RGB and RGB-D approaches.

1. Introduction

Salient object detection (SOD) aims to identify regions in a scene that visually attract human attention most [23,33, 44]. Recently, this fundamental task plays an important role in various computer vision applications [15,21,29,37], e.g., visual tracking, image segmentation and object recognition.

In the past, most saliency methods [11,28,32,34,41,50] focus on extracting hand-crafted features based on limited domain-specific knowledge, which may limit their generalization ability in different scenarios. Recently, CNNs-based methods have yielded a qualitative leap in performances due to the powerful ability of CNNs [25] to hierarchically extract informative features. Many works [10,22,30,44,45,47] focus on identifying saliency regions based on RGB images and have achieved superior performance, yet they still remain challenging in some complex scenarios, such as similar foreground and background, low-intensity environment. Depth information containing various depth cues such as spatial structure and 3D layout has been demonstrated to alleviate those issues in SOD [3, 4, 35]. In this paper, we mainly focus on effectively using RGB-D data to enhance model's robustness especially in challenging scenes. As exemplified in Fig. 1, RGB-D methods are superior to RGB methods in terms of a complex scene, in which the salient object shares similar appearance with its surroundings.

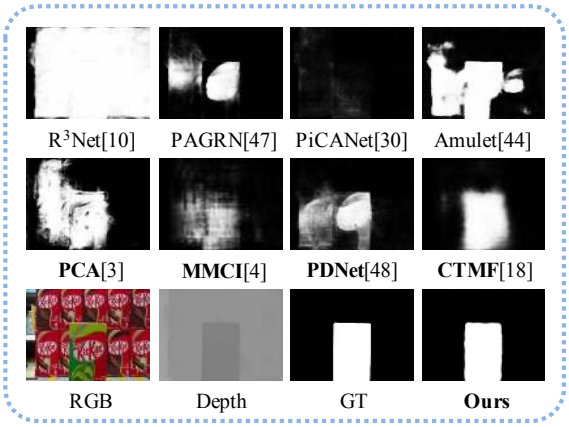

Figure 1. Saliency maps of several state-of-the-art CNNs-based methods in a complex scene. RGB-D methods are marked in **bold**.

Nonetheless, previous works for RGB-D SOD share some common limitations: 1) Most CNNs-based methods [4, 18, 35] generally fuse RGB and depth features by direct concatenation or summation at a shallow or deep stage. The complementarity of multi-level RGB and depth information is not taken into account. Specifically, the deep features can provide discriminative semantic information while the shallow features also contain affluent local details for accurately identifying salient objects. A recent work [3] concentrates on fusing multi-level information for prediction and achieves better performance. 2) Multiple objects in a scene have large variations in both depth and scale. Explor-

*Prof.Zhang is the corresponding author.

ing the relationship between depth cues and objects with different scales can further provide vital guidance cues for accurately locating salient regions. However, to our best knowledge, this relevance has never been researched in previous SOD works. 3) Studies show that people perceive visual information using an Internal Generative Mechanism (IGM) [17, 46]. In the IGM, saliency captured by human is not a straight translation of the ocular input, but a result of a series of active inferences of brains, especially in complex scenes. However, the benefits of IGM for comprehensively understanding a scene and capturing accurate saliency regions have never been explored in previous works. Particularly, the fused feature is directly used for prediction while the internal semantic relation in the fused feature is ignored.

In order to address the aforementioned limitations, we propose a depth-induced multi-scale recurrent attention network (DMRANet) for saliency detection as illustrated in Fig. 2. There are three main contributions of our DMRANet. **First**, we design an effective depth refinement block (DRB) taking advantages of residual connections to fully extract and fuse complementary RGB and depth features in multiple levels. **Second**, we innovatively design a depth-induced multi-scale weighting (DMSW) module. In this module, the relationship between depth information and objects with different scales is explored for the first time in saliency detection task (see Fig. 4). Ablation analysis shows that utilizing this relevance can improve detection accuracy and facilitate the integration of RGB and depth data. After the two procedures, a fused feature with abundant saliency cues is generated. **Third**, we design a novel recurrent attention module (RAM) inspired by the IGM of human brain. Our RAM can iteratively generate more accurate saliency results in a coarse-to-fine manner by comprehensively learning the internal semantic relation of the fused feature. Specifically, when inferring the current result, our RAM retrieves the previous memory to aid current decision. This can progressively optimize local details with memory-oriented scene understanding for generating the final optimal saliency result. This module boosts our model's performance by a large margin. **In addition**, we also create a large scale RGB-D dataset with 1200 paired images containing more complex scenarios, such as multiple or transparent objects, similar foreground and background, complex background, low-intensity environment. This challenging dataset can comprehensively evaluate saliency models and contribute to further studies in saliency field.

Furthermore, extensive experiments on seven datasets demonstrate that our method achieves consistently superior performance over 16 state-of-the-art 2D and 3D approaches. The code and results can be found at `https://github.com/OIPLab-DUT/DMRA_RGBD-SOD`. Moreover, to facilitate research in this field, all those partitioned datasets we collected are shared in a ready-to-use manner.

2. Related work

RGB-D saliency detection. Although many works [10, 14, 22, 30, 44, 45, 47] have devoted to RGB saliency detection and have achieved appealing performance, they might fail when coping with complex scenarios, such as multiple or transparent objects, similar foreground and background, complex background and low-intensity environment. Depth cues with affluent spatial structure and 3D layout information can contribute to handling those cases [3, 8, 11, 18, 32]. In our work, we mainly focus on RGB-D saliency detection and intend to improve detector's performance in complex scenes.

Previous RGB-D saliency detection methods can be generally classified into two categories: (1) manually designing hand-crafted features; (2) automatically extracting features with CNNs. For the first category, [32] utilize a multi-stage model combining RGB-produced saliency with new depth-induced saliency for SOD. [16, 24] present saliency methods based on anisotropic center-surround difference or local background enclosure. [36] exploit the normalized depth prior and the global-context prior for SOD. Those methods, mainly relying on hand-crafted features and lacking of high-level representations, are unadapted for understanding global context. Recently, CNNs have significantly pushed the performance of vision tasks for its powerful ability in hierarchically extracting informative features. [35] use hand-crafted features to train a CNN-based model and achieve significant improvements over traditional methods. [4, 18] utilize two-stream CNNs-based models but perform fusion by directly concatenating or adding paired features at shallow or deep layers. [48] propose a prior-model guided depth-enhanced network for SOD. Those fusion strategies do not take full advantage of multi-level complementary cues. A recent work [3] designs a fusion network, in which cross-level features are progressively combined, and achieves better performance. Besides, we observe that some schemes [4, 18, 48] adopt extra pre-training or post-processing operations for improving model's performance, which entangles the training process to some extent, whereas our network is trained in an end-to-end manner.

3. The proposed method

We first describe the overall architecture briefly in Sec. 3.1. Then, we discuss our multi-level fusion strategy and its key component-DRB in Sec. 3.2 and give a detailed depiction of our DMSW module in Sec. 3.3. Finally, we elaborate on the RAM which significantly improves the performance in Sec. 3.4.

3.1. The overall architecture

Our network architecture, shown in Fig. 2, follows a two-stream model. The two streams have the same structure,

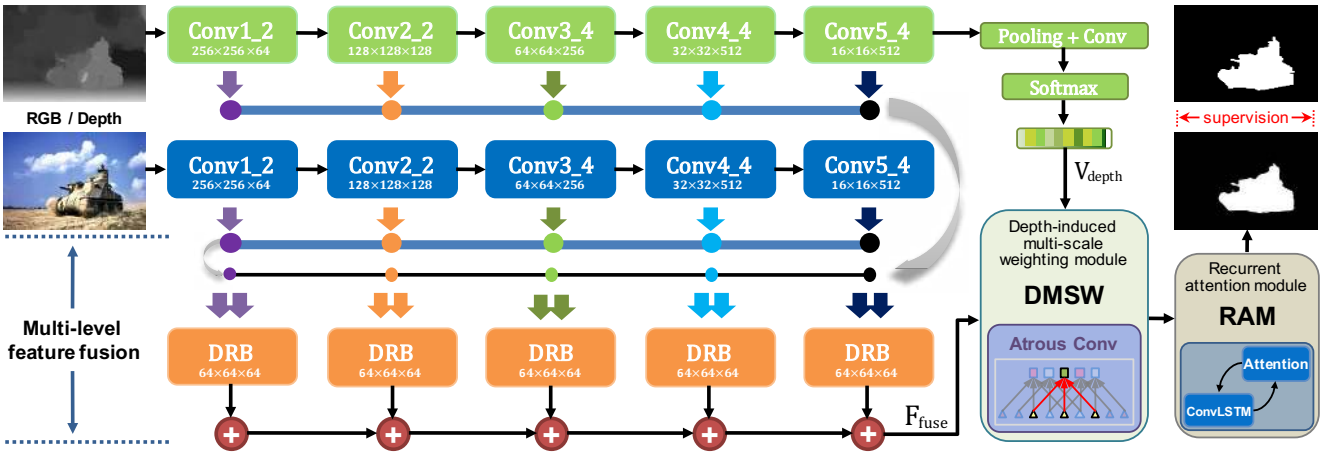

Figure 2. The overall architecture of our DMRANet.

where 5 convolutional blocks of VGG-19 [39] are maintained and the last pooling and fully-connected layers are discarded for making a better fit with our task. The only difference between two streams is that the depth stream is further processed to learn a depth vector. We refine and fuse paired side-out features in multiple layers by employing the proposed DRB. Then, the depth vector and the fused feature are fed into a DMSW module, in which multi-scale features generated from the fused feature are integrated based on the guidance from the depth vector. Moreover, we boost our model's performance by a novel RAM which ably combines the attention mechanism and ConvLSTM [38]. Finally, the saliency maps are supervised by the ground truths. Our network is trained in an end-to-end manner.

3.2. Multi-level Fusion Module

Considering the complementarity between paired depth and RGB cues in multiple layers, we design a simple yet effective DRB using residual connections [20] to fully extract and fuse multi-level paired complementary information.

Depth refinement block. As illustrated in Fig. 3, the inputs f_i^{RGB} and f_i^{depth} represent the side-out features from the RGB and depth streams in the i-th level respectively. We feed f_i^{depth} into a series of weight layers $\Psi(\cdot)$ containing two convolutional layers and two PReLU activation functions [19] to learn a depth residual $\Delta depth_i = \Psi(f_i^{depth})$. Then, the depth residual is added to the RGB feature by residual connection to learn a fused feature $f_i^{fuse} = f_i^{RGB} + \Delta depth_i$. In this way, complementary clues in the i-th level are fused effectively. Then, we reshape (i.e., up-sample with bilinear interpolation or down-sample with max-pooling operation) f_i^{fuse} to the same resolution. A conventional residual unit [20] $\Re(\cdot)$ is followed for re-scaling feature values and then a 1×1 convolution operation W_i is used to adjust the channel dimension. The final feature in the i-th level is defined as $f_i = W_i * \Re(reshape(f_i^{fuse}))$, which is 1/4 of the input

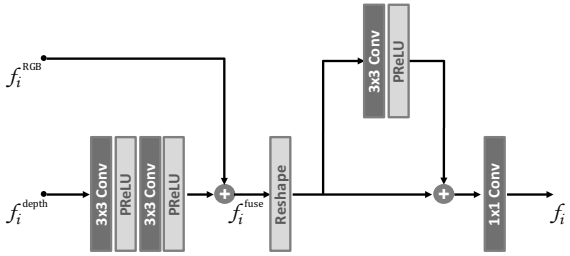

Figure 3. Detailed diagram of Depth Refinement Block (DRB).

spatial resolution with 64 channels. Finally, all features f_i in multiple layers are summated as $F_{fuse} = \sum_{i=1}^{N} f_i$ in an element-wise manner, where N=5 denotes the total number of convolutional blocks. In this way, discriminative multi-level RGB and depth features are effectively learned and fused. This fusion strategy enables our model to produce more accurate saliency results because of the comprehensive combination of both local spatial details and global semantic information.

3.3. Depth-induced Multi-scale Weighting Module

Considering that an image consists of multiple distinct objects with different sizes, scales and laid across different spatial locations in numerous layouts, we propose a depth-induced multi-scale weighting (DMSW) module. In this module, depth cues are further connected with multi-scale features to accurately locate salient objects.

As shown in Fig. 4, depth cues with abundant spatial information are further processed to learn a depth vector to guide the weight allocation of multi-scale features. To be specific, in order to capture multi-scale context features, we impose a global pooling layer and several parallel convolutional layers with different kernel sizes and different dilation rates on the input feature F_{fuse}. In this way, six multi-scale features F_m (m = 1, 2, . . . , 6) with the same resolution but different contexts are generated. Detailed parameters are shown in Fig. 4. Compared with classic convolution operation, dilated convolution can increase the size of the

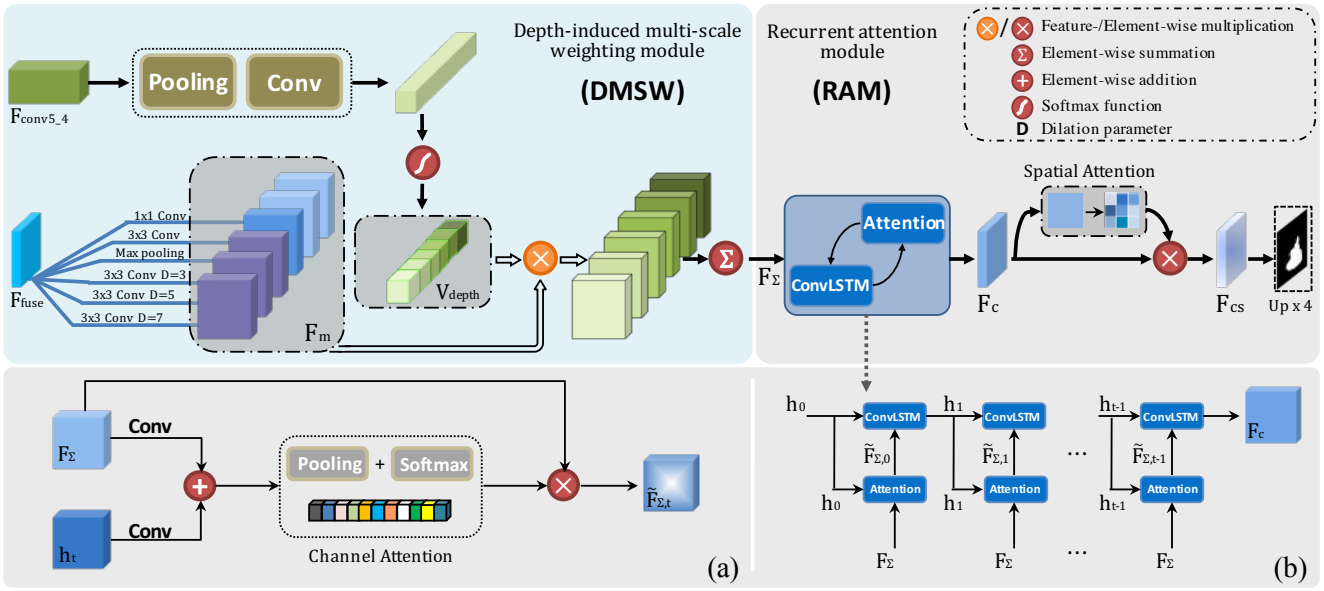

Figure 4. Detailed diagram of DMSW and RAM sub-modules. In RAM, (b) is the details of RAM and (a) is the details of attention block.

receptive field without sacrificing image resolution and redundant computation [5–7, 43]. Meanwhile, in order to obtain the corresponding depth vector, a global average pooling layer and a convolutional layer are imposed on F_{conv5_4} in the depth stream. Then we use a softmax function δ to obtain the depth vector $V_{depth} \in \mathbb{R}^{1\times1\times M}$, which can act as the scale factor for weighting each multi-scale feature F_m, where M responds to the maximum of m. Finally, all multi-scale features F_m are weighted based on depth vector V_{depth} and then summated to form the final output F_Σ. Formally, the DMSW module can be defined as:

$$V_{depth} = \delta(W_b * AvgPooling(F_{conv5\_4})), \quad (1)$$

$$F_m = \xi(F_{fuse}; \theta_m), \quad (2)$$

$$F_\Sigma = \sum_{m=1}^{M} V_{depth}^m \times F_m, \quad (3)$$

where $*$ and W_b denote convolution operation and corresponding parameters. $\delta(\cdot)$ represents the softmax function. $\xi(\cdot)$ denotes those parallel convolution or pooling operations and θ_m is the parameters to be learned in the m-th branch. V_{depth}^m represents the weight of the corresponding multi-scale feature F_m and $\times$ means the feature-wise multiplication.

In summary, it is beneficial to introduce depth cues to learn the contribution of multi-scale features for determination of salient objects especially when objects of different sizes appear at different depths. This module can also be regarded as a deeper fusion of RGB and depth information.

3.4. Recurrent Attention Module

Note that our model has outperformed all other state-of-the-art methods almost across all datasets by directly using the fused feature F_Σ for prediction, as described in ablation analysis. However, we claim that exploring the semantic relation inside the fused feature is essential, motivated by the Internal Generative Mechanism (IGM) [17] in human visual system. In this section, we boost our model's performance by a novel recurrent attention module (RAM). This module, drawing core ideas from the IGM, can comprehensively understand a scene and learn the internal semantic relation of the fused feature. To be specific, in order to infer conspicuous objects, the IGM recurrently deduces and predicts saliency based on memory stored in the brain, while uncertain information that is not important will be discarded.

Inspired by the IGM, we propose the RAM by ably combining attention mechanism and ConvLSTM [38]. In this way, the RAM can retrieve the previous memory to aid current decision when inferring the current result. It iteratively learns the spatio-temporal dependencies between different semantics and progressively optimizes detection details with memory-oriented scene understanding. Concretely, for the attention block (see Fig. 4(a)), h_t stands for the previous memory for scene understanding and F_Σ is the input feature. The subscript t denotes time steps in ConvLSTM. Both h_t and F_Σ are followed by a convolutional layer and then we merge the output features by element-wise summation. Then, a global average pooling and a softmax function are used to generate the channel-wise attention map $Att_c(h_t, F_\Sigma) \in \mathbb{R}^{1\times1\times C}$, in which C denotes the number of channels of F_Σ. By performing element-wise multiplication on $Att_c(h_t, F_\Sigma)$ and F_Σ, a more informative feature $\widetilde{F}_{\Sigma,t}$ is produced. This procedure can be defined as:

$$Att_c(h_t, F_\Sigma) = \delta(AvgPooling(W_0 * h_t + W_1 * F_\Sigma)), \quad (4)$$

$$\widetilde{F}_{\Sigma,t} = Att_c(h_t, F_\Sigma) \otimes F_\Sigma, \quad (5)$$

where W_* are convolution parameters. $\otimes$ means element-wise multiplication. Next, in Fig. 4(b), $\widetilde{F}_{\Sigma,t}$ is fed into ConvLSTM to further learn the spatial correlation between different semantic features. The ConvLSTM is calculated by

$$\begin{aligned} i_t &= \sigma(W_{xi} * \widetilde{F}_{\Sigma,t} + W_{hi} * h_{t-1} + W_{ci} \circ c_{t-1} + b_i), \\ f_t &= \sigma(W_{xf} * \widetilde{F}_{\Sigma,t} + W_{hf} * h_{t-1} + W_{cf} \circ c_{t-1} + b_f), \\ c_t &= f_t \circ c_{t-1} + i_t \circ \tanh(W_{xc} * \widetilde{F}_{\Sigma,t} + W_{hc} * h_{t-1} + b_c), \\ o_t &= \sigma(W_{xo} * \widetilde{F}_{\Sigma,t} + W_{ho} * h_{t-1} + W_{co} \circ c_{t-1} + b_o), \\ h_t &= o_t \circ \tanh(c_t), \end{aligned} \tag{6}$$

where $\circ$ denotes the Hadamard product and $\sigma(\cdot)$ is sigmoid function. i_t, f_t and o_t stand for input, forget and output gates, respectively. c_t stores the earlier information. All W_* and b_* are model parameters to be learned. h_0 and c_0 are initialized to 0. After N steps, where we set $N = 3$ in this work, a channel-refined feature $F_c = h_N$ is generated.

In addition, we add a common spatial attention block to emphasize the contribution of each pixel for the final saliency prediction. We first learn a spatial-wise attention map $Att_s(F_c) = \sigma(W_s * F_c)$, where $*$ and W_s represent a 1×1 convolution operation and corresponding parameters, respectively. Then $Att_s(F_c) \in \mathbb{R}^{W \times H \times 1}$ and F_c are multiplied in an element-wise manner to get a spatial weighted feature $F_{cs} = Att_s(F_c) \otimes F_c$.

Eventually, F_{cs} is followed by a 1×1 convolution layer and up-sample operation to get the final saliency map S_{map}.

4. Experiments

4.1. Dataset

We evaluate the effectiveness of our network on our proposed dataset and other six public datasets.

NJUD [24]: contains 1985 images (the latest version), which are collected from the Internet, 3D movies and photographs taken by a Fuji W3 stereo camera. **NLPR [32]:** includes 1000 images captured by Kinect. **LFSD [27]:** contains 100 images captured by Lytro camera. **STEREO [31]:** contains 797 stereoscopic images downloaded from the Internet. **RGBD135 [8]:** contains 135 images captured by Kinect. **SSD [26]:** contains 80 images picked up from three stereo movies.

Ours: Compared to other datasets, ours is more challenging containing many complex scenes (e.g., multiple or transparent objects, similar foreground and background, complex background and low-intensity environment). The bottom five rows marked in Fig. 5 show some representative scenes in our dataset. Our dataset contains 800 indoor and 400 outdoor scenes paired with corresponding depth maps [40] and ground truths. This challenging dataset can contribute to comprehensively evaluating saliency models. More details about this dataset can de found at the github page.

Training and test: Our dataset is randomly divided into two parts: 800 images for training and the rest 400 for testing. For other datasets, we adopt the same splitting way as [3, 4, 18] to guarantee a fair comparison. We split 1485 samples from NJUD and 700 samples from NLPR for training. The remaining images in these two datasets and other four datasets are all for testing to verify the generalization ability of saliency models. To prevent overfitting, we augment the training set by flipping, cropping and rotating.

4.2. Experimental setup

Evaluation metrics. For comprehensively evaluating various methods, we adopt five evaluation metrics including precision-recall (PR) curve, F-measure (F_β) [1], mean absolute error (MAE) [2] and recently proposed S-measure (S_λ) [12] and E-measure (E_γ) [13]. Concretely, saliency maps are binarized using a series of thresholds and then pairs of precision and recall are computed to plot the PR curve. The F-measure can evaluate the overall performance. The MAE represents the average absolute difference between the saliency map and ground truth. The S-measure can evaluate the spatial structure similarities and the E-measure can jointly capture image level statistics and local pixel matching information. For MAE, lower value is better and for others, higher is better.

Implementation details. Our method is implemented with pytorch toolbox and trained on a PC with GTX 1080 GPU and 16 GB memory. The input image is uniformly resized to 256×256. The momentum, weight decay and learning rate are set as 0.9, 0.0005 and 1e-10, respectively. During training, we use softmax entropy loss and the network converges after 50 epochs with mini-batch size 2.

4.3. Comparison with state-of-the-art

We compare our method with 16 state-of-the-art ones, including 5 latest CNNs-based RGB-D methods: PCA [3], PDNet [48], MMCI [4], CTMF [18], DF [35]; 5 traditional RGB-D methods: MB [49], CDCP [50], NLPR [32], DES [8], DCMC [9]; 6 top ranking CNNs-based RGB methods: PiCANet [30], PAGRN [47], R^3Net [10], Amulet [44], UCF [45], DSS [22]. For fair comparisons, we use the released code and their default parameters to reproduce those methods. In terms of methods without released source code, we use their published results for comparisons.

Quantitative Evaluation. Tab. 1 and Tab. 2 show the validation results in terms of four evaluation metrics on seven datasets. We can see that our model achieves significant outperformance over all other methods. The PR curves in Fig. 6 also consistently demonstrate the superior performance of our method. Especially, ours outperforms all other methods by a dramatic margin on our proposed dataset, NLPR and STEREO, where the images are comparably complicated. It further indicates that our model is

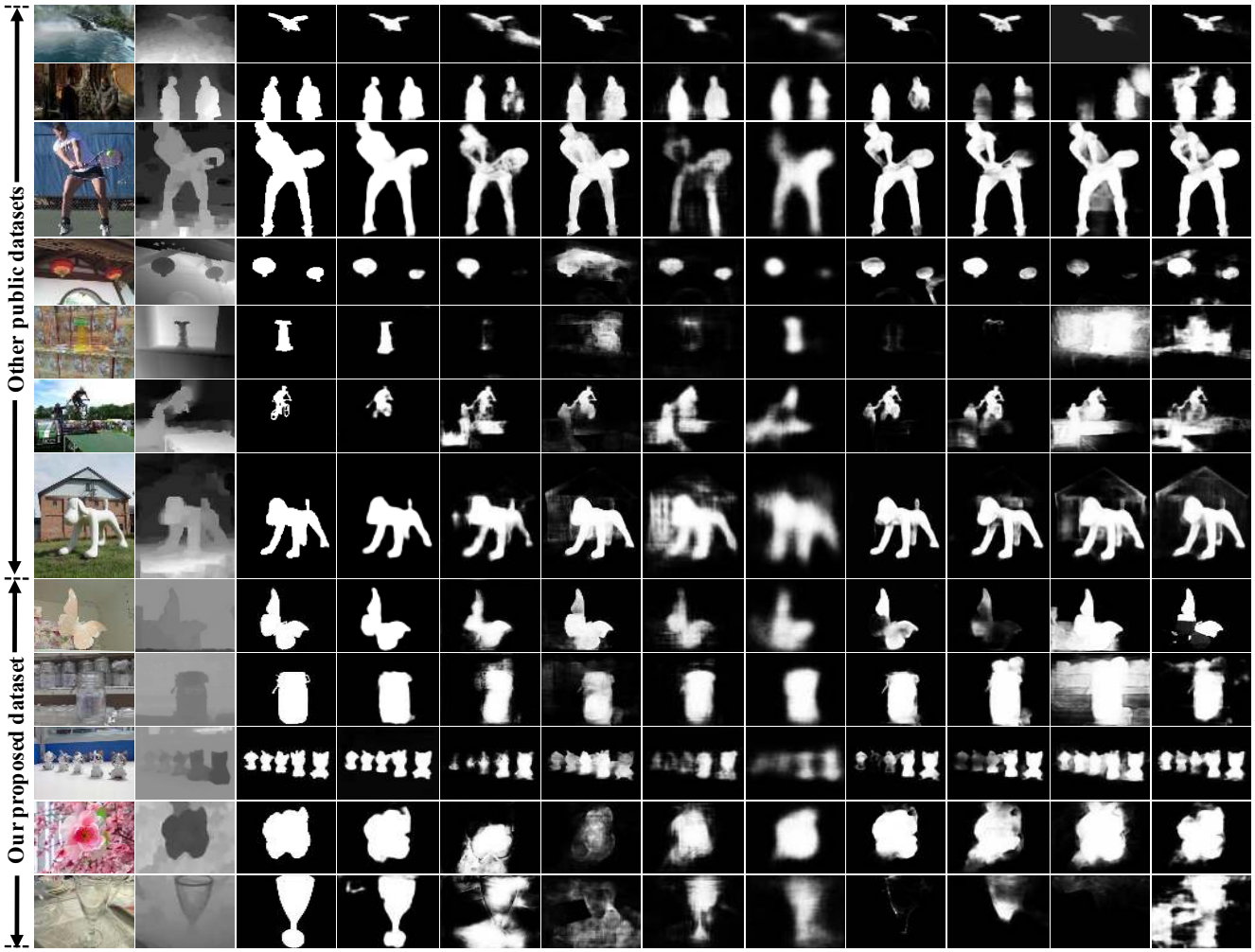

Figure 5. Comparisons of ours with state-of-the-art CNNs-based methods. Those methods are top ranking ones in quantitative evaluation. Obviously, our results are more consistent with the ground truths ('GT'), especially in complex scenes, such as cluttered background (5^{th} and 6^{th} rows), low-contrast (11^{th} row), transparent object (9^{th} and 12^{th} rows) as well as multiple and small objects (10^{th} row).

more powerful in dealing with the complex scenes.

Qualitative Evaluation. We also visually compare our method with the most representative methods as shown in Fig. 5. From those results, we can observe that our saliency maps are closer to the ground truths. For example, other methods are difficult to distinguish salient objects in complex environments (see the 5^{th} and 6^{th} rows), while ours can precisely identify the whole object. And our DMRANet can more accurately locate and detect the entire conspicuous objects with sharp details than others in more challenging scenes such as low-contrast, transparent object as well as multiple and small objects (see the 9^{th}-12^{th} rows). Those results further verify the effectiveness and robustness of our proposed DMRANet.

4.4. Ablation analysis

In this section, we perform ablation analysis over each component of the DMRANet and further investigate their relative importance and specific contribution.

Performance of DRB. In order to verify the effectiveness of the proposed multi-level fusion strategy, we evaluate the performance of a common fusion strategy (see Fig. 7 (a)) and our DRB fusion strategy (denoted as 'Baseline' and '+DRB', respectively). As shown in Tab. 3 and Fig. 8, '+DRB' consistently outperforms 'Baseline' across all datasets. The predictions produced by our DRB contain more local details than 'Baseline' in Fig. 9. This advance further confirms the superiority of our DRB in effectively and abundantly extracting and fusing multi-level paired complementary information.

Performance of DMSW module. One of our core claims is that incorporating depth cues with multi-scale features can help locate saliency regions. To give evidence for this claim, we add the DMSW module ('+DMSW') to previous '+DRB' model. Results in Tab. 3 and Fig. 8 show that our DMSW module achieves impressive accuracy gains on

*	Ours				NJUD				NLPR				STEREO	
	E_γ	S_λ	F_β	MAE	E_γ	S_λ	F_β	MAE	E_γ	S_λ	F_β	MAE	E_γ	S_λ
Ours	**0.927**	**0.888**	**0.883**	**0.048**	**0.908**	**0.886**	**0.872**	**0.051**	**0.942**	**0.899**	**0.855**	**0.031**	**0.920**	**0.886**
PCA	0.858	0.801	0.760	0.100	0.896	0.877	0.844	0.059	0.916	0.873	0.794	0.044	0.905	0.880
PDNet	0.861	0.799	0.757	0.112	0.890	0.883	0.832	0.062	0.876	0.835	0.740	0.064	0.903	0.874
MMCI	0.855	0.791	0.753	0.113	0.878	0.859	0.813	0.079	0.871	0.855	0.729	0.059	0.890	0.856
CTMF	0.884	0.834	0.792	0.097	0.864	0.849	0.788	0.085	0.869	0.860	0.723	0.056	0.870	0.853
DF	0.842	0.730	0.748	0.145	0.818	0.735	0.744	0.151	0.838	0.769	0.682	0.099	0.844	0.763
PiCANet	0.895	0.832	0.826	0.080	0.880	0.847	0.806	0.071	0.895	0.834	0.761	0.053	0.904	0.868
PAGRN	0.883	0.831	0.836	0.079	0.882	0.829	0.827	0.081	0.907	0.844	0.795	0.051	0.900	0.851
R^3Net	0.833	0.819	0.781	0.113	0.838	0.837	0.775	0.092	0.788	0.798	0.649	0.101	0.856	0.855
Amulet	0.880	0.846	0.803	0.083	0.859	0.843	0.798	0.085	0.852	0.848	0.722	0.062	0.897	0.881
UCF	0.848	0.833	0.766	0.108	0.830	0.829	0.758	0.109	0.835	0.837	0.701	0.082	0.874	0.867
DSS	0.831	0.767	0.732	0.127	0.853	0.807	0.776	0.108	0.879	0.816	0.755	0.076	0.885	0.841
MB	0.691	0.607	0.577	0.156	0.643	0.534	0.492	0.202	0.814	0.714	0.637	0.089	0.693	0.579
CDCP	0.794	0.687	0.633	0.159	0.751	0.673	0.618	0.181	0.785	0.724	0.591	0.114	0.801	0.727
NLPR	0.767	0.568	0.659	0.174	0.722	0.530	0.625	0.201	0.772	0.591	0.520	0.119	0.781	0.567
DES	0.733	0.659	0.668	0.280	0.421	0.413	0.165	0.448	0.735	0.582	0.583	0.301	0.451	0.473
DCMC	0.712	0.499	0.406	0.243	0.796	0.703	0.715	0.167	0.684	0.550	0.328	0.196	0.838	0.745

Table 1. Quantitative comparison of E-measure, S-measure, F-measure and MAE on our proposed dataset and six widely-used RGB-D datasets. The best three results are shown in **boldface**, red, and green fonts respectively. Our method ranks first on all datasets and evaluation metrics. From top to bottom: CNNs-based RGB-D methods, the latest RGB methods and traditional RGB-D methods.

*	STEREO		LFSD				RGBD135				SSD			
	F_β	MAE	E_γ	S_λ	F_β	MAE	E_γ	S_λ	F_β	MAE	E_γ	S_λ	F_β	MAE
Ours	**0.868**	**0.047**	**0.899**	**0.847**	**0.849**	**0.075**	**0.945**	**0.901**	**0.857**	**0.029**	**0.892**	**0.857**	**0.821**	**0.058**
PCA	0.845	0.061	0.846	0.800	0.794	0.112	0.909	0.845	0.763	0.049	0.883	0.843	0.786	0.064
PDNet	0.833	0.064	0.872	0.845	0.824	0.109	0.915	0.868	0.800	0.050	0.813	0.802	0.716	0.115
MMCI	0.812	0.080	0.840	0.787	0.779	0.132	0.899	0.847	0.750	0.064	0.860	0.814	0.748	0.082
CTMF	0.786	0.087	0.851	0.796	0.781	0.120	0.907	0.863	0.765	0.055	0.837	0.776	0.709	0.100
DF	0.761	0.142	0.801	0.685	0.566	0.130	0.801	0.685	0.566	0.130	0.802	0.742	0.709	0.151
PiCANet	0.835	0.062	0.806	0.761	0.730	0.134	0.928	0.854	0.797	0.042	0.882	0.832	0.775	0.068
PAGRN	0.856	0.067	0.831	0.779	0.786	0.117	0.919	0.858	0.834	0.044	0.862	0.793	0.762	0.088
R^3Net	0.800	0.084	0.771	0.797	0.791	0.141	0.868	0.847	0.728	0.066	0.833	0.815	0.747	0.095
Amulet	0.842	0.062	0.863	0.827	0.817	0.101	0.866	0.842	0.725	0.070	0.843	0.828	0.756	0.087
UCF	0.808	0.083	0.816	0.811	0.773	0.138	0.854	0.835	0.717	0.089	0.807	0.795	0.693	0.117
DSS	0.814	0.087	0.778	0.718	0.694	0.166	0.855	0.763	0.697	0.098	0.834	0.786	0.752	0.116
MB	0.572	0.178	0.631	0.538	0.543	0.218	0.798	0.661	0.588	0.102	0.633	0.499	0.414	0.219
CDCP	0.680	0.149	0.737	0.658	0.634	0.199	0.806	0.706	0.583	0.119	0.714	0.604	0.524	0.219
NLPR	0.716	0.179	0.742	0.558	0.708	0.211	0.850	0.577	**0.857**	0.097	0.726	0.562	0.551	0.200
DES	0.223	0.417	0.475	0.440	0.228	0.415	0.786	0.627	0.689	0.289	0.383	0.341	0.073	0.500
DCMC	0.761	0.150	0.842	0.754	0.815	0.155	0.674	0.470	0.228	0.194	0.790	0.706	0.684	0.168

Table 2. Continuation of Table 1.

all datasets by comparing '+DMSW' and '+DRB'. From Fig. 9, we can see '+DMSW' can identify more saliency regions compared with '+DRB'. Those results demonstrate the advantage of our DMSW module in sufficiently utilizing depth cues and multi-scale information. Moreover, we also verify the benefits of utilizing the relationship between depth cues and multi-scale features by performing a new model, in which features at multiple scales are integrated by a 1×1 convolution operation instead of depth cues (denoted as '+DMSW (w/o d)'). Results in Tab. 3 and Fig. 8 show that removing depth guidance degrades performance to some extent. Those results also demonstrate that the combination of depth information and multi-scale features can improve the detection accuracy. In addition, it is important to note that our model has outperformed all other methods almost across all datasets at this stage. This fact further verifies the strength of our proposed module.

Performance of RAM. In this section, we evaluate the performance of our RAM. By comparing visual results in Fig. 9, we observe our RAM can further suppress background irritations and substantially optimize detection details. In addition, we replace the RAM with a basic channel-spatial attention block [42] (denoted as '+Att(common)') in Fig. 7 (b). Results in Tab. 3 suggest that our RAM is superior to '+Att(common)' and boosts model's performance by a large margin. We attribute this advance to its power-

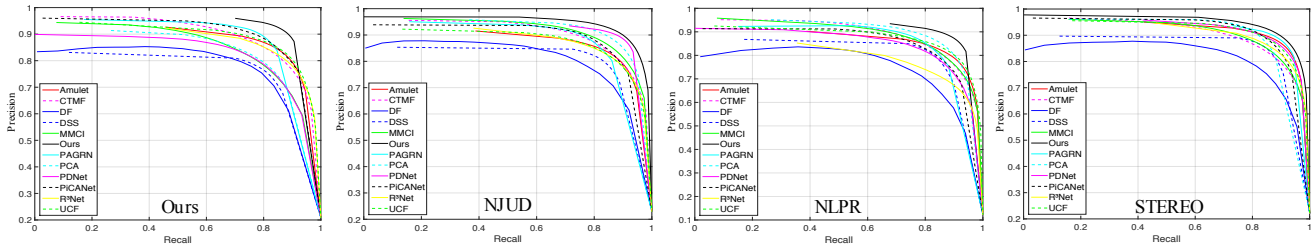

Figure 6. The PR curves of the proposed method and other state-of-the-art approaches across four datasets.

*	Ours		NJUD		NLPR		STEREO		LFSD		RGBD135		SSD	
	F_β	MAE	F_β	MAE	F_β	MAE	F_β	MAE	F_β	MAE	F_β	MAE	F_β	MAE
Baseline	0.828	0.070	0.820	0.068	0.758	0.051	0.822	0.067	0.822	0.094	0.780	0.047	0.758	0.081
+DRB	0.839	0.065	0.828	0.064	0.774	0.046	0.828	0.064	0.825	0.090	0.792	0.043	0.768	0.076
+DMSW(w/o d)	0.855	0.061	0.844	0.062	0.805	0.044	0.837	0.061	0.836	0.087	0.823	0.042	0.774	0.076
+DMSW	0.861	0.057	0.850	0.059	0.801	0.042	0.852	0.057	0.836	0.086	0.828	0.040	0.783	0.075
+Att(common)	0.869	0.054	0.860	0.055	0.827	0.036	0.859	0.053	0.847	0.081	0.842	0.032	0.809	0.064
+RAM(Ours)	**0.883**	**0.048**	**0.872**	**0.051**	**0.855**	**0.031**	**0.868**	**0.047**	**0.849**	**0.075**	**0.857**	**0.029**	**0.821**	**0.058**

Table 3. Ablation analysis on seven datasets. Obviously, each component of our DMRANet can provide additional accuracy gains.

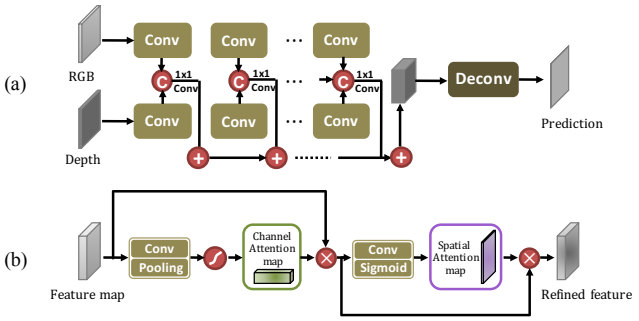

Figure 7. Diagrams of ablation analysis. (a) Baseline. 'C' means concatenation operation. (b) Att(common).

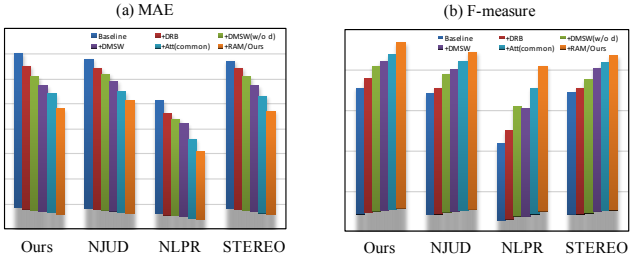

Figure 8. Histograms of F-measure and MAE on four datasets.

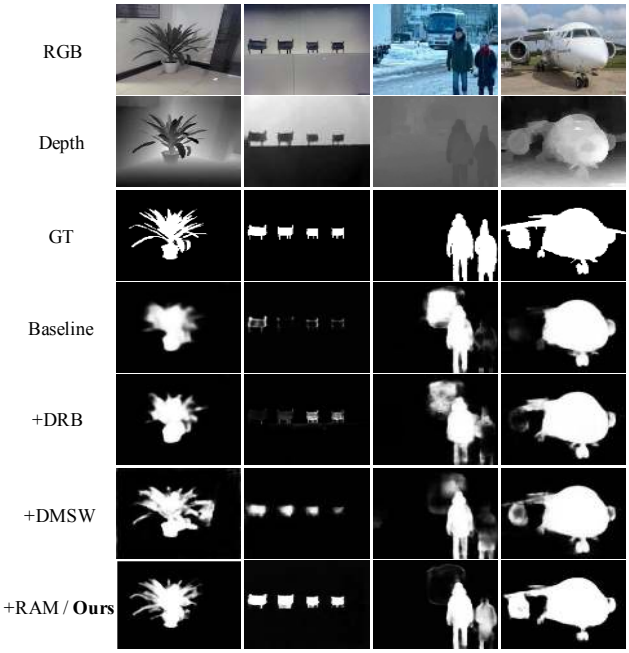

Figure 9. The visual results of ablation analysis.

ful ability in progressively optimizing detection details with memory-oriented scene understanding.

5. Conclusion

In this work, our proposed 'DMRANet' enhances the performance of saliency detection from three aspects: 1) fully extracts and fuses multi-level paired complementary features by using a simple yet effective DRB; 2) innovatively combines depth cues with multi-scale information to accurately locate and identify salient objects; 3) progressively generates more accurate saliency results through a novel recurrent attention model. In addition, we build a large scale RGB-D saliency dataset with 1200 paired images containing more challenging scenes. We comprehensively validate the effectiveness of each component of our network and show the accumulated accuracy gains gradually. Experiment results also demonstrate that our method achieves new state-of-the-art performance on seven RGB-D datasets.

Acknowledgment

This work was supported by the National Natural Science Foundation of China(61605022 and U1708263) and the Fundamental Research Funds for the Central Universities(DUT19JC58). The authors are grateful to the reviewers for their suggestions in improving the quality of the paper.

References

[1] Radhakrishna Achanta, Sheila S. Hemami, Francisco J. Estrada, and Sabine Süsstrunk. Frequency-tuned salient region detection. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1597–1604, 2009.

[2] Ali Borji, Dicky N. Sihite, and Laurent Itti. Salient object detection: a benchmark. In *European Conference on Computer Vision (ECCV)*, pages 414–429, 2012.

[3] Hao Chen and Youfu Li. Progressively complementarity-aware fusion network for rgb-d salient object detection. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3051–3060, 2018.

[4] Hao Chen, Youfu Li, and Dan Su. Multi-modal fusion network with multi-scale multi-path and cross-modal interactions for rgb-d salient object detection. *Pattern Recognition*, 86:376–385, 2019.

[5] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):834–848, 2018.

[6] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[7] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. *European Conference on Computer Vision (ECCV)*, pages 833–851, 2018.

[8] Yupeng Cheng, Huazhu Fu, Xingxing Wei, Jiangjian Xiao, and Xiaochun Cao. Depth enhanced saliency detection method. In *International Conference on Internet Multimedia Computing and Service (ICIMCS)*, pages 23–27, 2014.

[9] Runmin Cong, Jianjun Lei, Changqing Zhang, Qingming Huang, Xiaochun Cao, and Chunping Hou. Saliency detection for stereoscopic images based on depth confidence analysis and multiple cues fusion. *IEEE Signal Processing Letters*, 23(6):819–823, 2016.

[10] Zijun Deng, Xiaowei Hu, Lei Zhu, Xuemiao Xu, Jing Qin, Guoqiang Han, and Pheng-Ann Heng. R^3net: Recurrent residual refinement network for saliency detection. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pages 684–690, 2018.

[11] Karthik Desingh, Madhava Krishna K, Deepu Rajan, and C. V. Jawahar. Depth really matters: Improving visual salient region detection with depth. In *British Machine Vision Conference (BMVC)*, 2013.

[12] Deng-Ping Fan, Ming-Ming Cheng, Yun Liu, Tao Li, and Ali Borji. Structure-measure: A new way to evaluate foreground maps. In *International Conference on Computer Vision (ICCV)*, pages 4558–4567, 2017.

[13] Deng-Ping Fan, Cheng Gong, Yang Cao, Bo Ren, Ming-Ming Cheng, and Ali Borji. Enhanced-alignment measure for binary foreground map evaluation. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pages 698–704, 2018.

[14] Deng-Ping Fan, Jiang-Jiang Liu, Shang-Hua Gao, Qibin Hou, Ali Borji, and Ming-Ming Cheng. Salient objects in clutter: Bringing salient object detection to the foreground. In *European Conference on Computer Vision (ECCV)*, pages 1597–1604. Springer, 2018.

[15] Deng-Ping Fan, Wenguan Wang, Ming-Ming Cheng, and Jianbing Shen. Shifting more attention to video salient object detection. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8554–8564, 2019.

[16] David Feng, Nick Barnes, Shaodi You, and Chris McCarthy. Local background enclosure for rgb-d salient object detection. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2343–2350, 2016.

[17] Dashan Gao, Sunhyoung Han, and Nuno Vasconcelos. Discriminant saliency, the detection of suspicious coincidences, and applications to visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(6):989–1005, 2009.

[18] Junwei Han, Hao Chen, Nian Liu, Chenggang Yan, and Xuelong Li. Cnns-based rgb-d saliency detection via cross-view transfer and multiview fusion. *IEEE Transactions on Systems, Man, and Cybernetics*, 48(11):3171–3183, 2018.

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *International Conference on Computer Vision (ICCV)*, pages 1026–1034, 2015.

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[21] Seunghoon Hong, Tackgeun You, Suha Kwak, and Bohyung Han. Online tracking by learning discriminative saliency map with convolutional neural network. *International Conference on Machine Learning (ICML)*, pages 597–606, 2015.

[22] Qibin Hou, Ming-Ming Cheng, Xiaowei Hu, Ali Borji, Zhuowen Tu, and Philip H.S. Torr. Deeply supervised salient object detection with short connections. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5300–5309, 2017.

[23] Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259, 1998.

[24] Ran Ju, Ling Ge, Wenjing Geng, Tongwei Ren, and Gangshan Wu. Depth saliency based on anisotropic center-surround difference. In *International Conference on Image Processing (ICIP)*, pages 1115–1119, 2014.

[25] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. *Communications of The ACM*, 60(6):84–90, 2017.

[26] Ge Li and Chunbiao Zhu. A three-pathway psychobiological framework of salient object detection using stereoscopic technology. In *International Conference on Computer Vision Workshops (ICCVW)*, pages 3008–3014, 2017.

[27] Nianyi Li, Jinwei Ye, Yu Ji, Haibin Ling, and Jingyi Yu. Saliency detection on light field. In *Conference on Computer*

Vision and Pattern Recognition (CVPR), pages 2806–2813, 2014.

[28] Xiaohui Li, Huchuan Lu, Lihe Zhang, Xiang Ruan, and Ming-Hsuan Yang. Saliency detection via dense and sparse reconstruction. In *International Conference on Computer Vision (ICCV)*, pages 2976–2983, 2013.

[29] Yin Li, Xiaodi Hou, Christof Koch, James M. Rehg, and Alan L. Yuille. The secrets of salient object segmentation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 280–287, 2014.

[30] Nian Liu, Junwei Han, and Ming-Hsuan Yang. Picanet: Learning pixel-wise contextual attention for saliency detection. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3089–3098, 2018.

[31] Yuzhen Niu, Yujie Geng, Xueqing Li, and Feng Liu. Leveraging stereopsis for saliency analysis. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 454–461, 2012.

[32] Houwen Peng, Bing Li, Weihua Xiong, Weiming Hu, and Rongrong Ji. Rgbd salient object detection: A benchmark and algorithms. In *European Conference on Computer Vision (ECCV)*, pages 92–109, 2014.

[33] Yongri Piao, Zhengkun Rong, Miao Zhang, Xiao Li, and Huchuan Lu. Deep light-field-driven saliency detection from a single view. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2019.

[34] Yao Qin, Huchuan Lu, Yiqun Xu, and He Wang. Saliency detection via cellular automata. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 110–119, 2015.

[35] Liangqiong Qu, Shengfeng He, Jiawei Zhang, Jiandong Tian, Yandong Tang, and Qingxiong Yang. Rgbd salient object detection via deep fusion. *IEEE Transactions on Image Processing*, 26(5):2274–2285, 2017.

[36] Jianqiang Ren, Xiaojin Gong, Lu Yu, Wenhui Zhou, and Michael Ying Yang. Exploiting global priors for rgb-d saliency detection. In *Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 25–32, 2015.

[37] Zhixiang Ren, Shenghua Gao, Liang-Tien Chia, and Ivor Wai-Hung Tsang. Region-based saliency detection and its application in object recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 24(5):769–779, 2014.

[38] Xingjian Shi, Zhourong Chen, Hao Wang, Dit Yan Yeung, Wai Kin Wong, and Wangchun Woo. Convolutional lstm network: a machine learning approach for precipitation nowcasting. *Neural Information Processing Systems (NIPS)*, pages 802–810, 2015.

[39] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *International Conference on Learning Representations (ICLR)*, 2015.

[40] Michael W. Tao, Sunil Hadap, Jitendra Malik, and Ravi Ramamoorthi. Depth from combining defocus and correspondence using light-field cameras. In *International Conference on Computer Vision (ICCV)*, pages 673–680, 2013.

[41] Wei-Chih Tu, Shengfeng He, Qingxiong Yang, and Shao-Yi Chien. Real-time salient object detection with a minimum spanning tree. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2334–2342, 2016.

[42] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. *European Conference on Computer Vision (ECCV)*, pages 3–19, 2018.

[43] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *International Conference on Learning Representations (ICLR)*, 2016.

[44] Pingping Zhang, Dong Wang, Huchuan Lu, Hongyu Wang, and Xiang Ruan. Amulet: Aggregating multi-level convolutional features for salient object detection. In *International Conference on Computer Vision (ICCV)*, pages 202–211, 2017.

[45] Pingping Zhang, Dong Wang, Huchuan Lu, Hongyu Wang, and Baocai Yin. Learning uncertain convolutional features for accurate saliency detection. In *International Conference on Computer Vision (ICCV)*, pages 212–221, 2017.

[46] Xiaoli Zhang, Xiongfei Li, Yuncong Feng, Haoyu Zhao, and Zhaojun Liu. Image fusion with internal generative mechanism. *Expert Systems With Applications*, 42(5):2382–2391, 2015.

[47] Xiaoning Zhang, Tiantian Wang, Jinqing Qi, Huchuan Lu, and Gang Wang. Progressive attention guided recurrent network for salient object detection. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 714–722, 2018.

[48] Chunbiao Zhu, Xing Cai, Kan Huang, Thomas H Li, and Ge Li. Pdnet: Prior-model guided depth-enhanced network for salient object detection. *arXiv preprint arXiv:1803.08636*, 2018.

[49] Chunbiao Zhu, Ge Li, Xiaoqiang Guo, Wenmin Wang, and Ronggang Wang. A multilayer backpropagation saliency detection algorithm based on depth mining. In *International Conference on Computer Analysis of Images and Patterns (CAIP)*, pages 14–23, 2017.

[50] Chunbiao Zhu, Ge Li, Wenmin Wang, and Ronggang Wang. An innovative salient object detection using center-dark channel prior. In *International Conference on Computer Vision Workshops (ICCVW)*, pages 1509–1515, 2017.